Robust identification in large scale random variables networks

Valery Kalyagin

LATNA market network analysis team
Laboratory of Algorithms and Technologies for Network Analysis (LATNA)
National Research University Higher School of Economics,
Nizhny Novgorod, Russia
vkalyagin@hse.ru

HSE, October 23, 2015

Outline

Random variables network

- Nodes are random variables.
- Weights of edges are given by some measure of association (similarity, dependance, ...).

Random variable network is a pair (X, γ):

- $X = (X_1, \ldots, X_N)-$random vector,
- $\gamma-$measure of association.

Network structures identification problem: identify a network structure (subgraph) by observations.
We consider the threshold graph identification problem.
Motivation:

- identification of the market graph in market network.
- model selection in Gaussian graphical model.

Threshold graph

- Random variable network $(X, \gamma) : X = (X_1, \dots, X_N), \gamma-$measure of association.
- Threshold graph (TG) is constructed by removing all edges with $\gamma_{i,j} := \gamma(X_i, X_j) \leq \gamma_0$ (γ_0- threshold). $\gamma_{i,j}$ - measure of association between nodes i and j.
- Popular network:=Pearson network: $\gamma^P_{i,j} = \rho_{i,j} = \frac{E(X_i - E(X_i))(X_j - E(X_j))}{\sigma_i \sigma_j}$

Threshold graph identification problem

Let $X(t), t = 1, 2, \ldots, n$ be a sample from the distribution of the random vector X. $X(t) = (X_1(t), \ldots, X_N(t))$

Problem: for a given threshold γ_0 identify the threshold graph from observations $X(t), t = 1, \ldots, n$.

Identification statistical procedure: map from the sample space $R^{N \times n}$ to the decision space $\mathcal{G}$, where
$\mathcal{G}$ - set of $N \times N$ symmetric matrices $G = (g_{i,j}); g_{i,j} \in \{0, 1\}$,
$i, j = 1, 2, \ldots, N$, $g_{i,i} = 0$, $i = 1, 2, \ldots, N$.
$G \in \mathcal{G}$ - adjacency matrices of all simple undirected graphs with N vertices. Total number of matrices in $\mathcal{G}$ is $L = 2^M$ with $M = N(N-1)/2$.
This is a multiple decision problem. Possible solution - multiple testing statistical procedures

Multiple testing statistical procedures.

Individual edge hypotheses:

$$h_{ij} : \gamma_{ij} \leq \gamma_0 \text{ vs } k_{ij} : \gamma_{ij} > \gamma_0.$$

Individual tests:

$$\varphi_{ij}(X) = \begin{cases} 1, & t_{ij}(X) > c_{ij} \\ 0, & t_{ij}(X) \leq c_{ij} \end{cases}$$

Multiple testing statistical procedure: statistical procedure, based on statistics of individual tests.

- Single step procedures (Bonferroni and others)
- Stepwise procedures (Holm, Hochberg and their modifications)

Pearson network

- Individual hypotheses (Pearson measure):
  $h_{ij} : \rho_{i,j} \leq \rho_0$ vs $k_{ij} : \rho_{i,j} > \rho_0$;
- $\varphi^P_{i,j}(x_i, x_j) = \begin{cases} 1, & z_{i,j} > c_{i,j} \\ 0, & z_{i,j} \leq c_{i,j} \end{cases}$
  where $z_{i,j} = \sqrt{n}\left(\frac{1}{2}\ln\left(\frac{1+r_{i,j}}{1-r_{i,j}}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right)$,
  $c_{i,j}$ is $(1-\alpha_{ij})$-quantile of standart normal distribution $N(0,1)$ $\alpha_{i,j}$ is the given significance level for individual edge i, j test,

$$r_{i,j} = \frac{\sum_{t=1}^{n}(x_i(t) - \overline{x_i})(x_j(t) - \overline{x_j})}{\sqrt{\sum_{t=1}^{n}(x_i(t) - \overline{x_i})^2 \sum_{t=1}^{n}(x_j(t) - \overline{x_j})^2}}$$

Pearson network

- Multiple testing single step (Bonferroni type) procedure:

$$\delta^P(x) = \begin{pmatrix} 1, & \varphi^P_{1,2}(x), & \dots, & \varphi^P_{1,N}(x) \\ \varphi^P_{2,1}(x), & 1, & \dots, & \varphi^P_{2,N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi^P_{N,1}(x), & \varphi^P_{N,2}(x), & \dots, & 1 \end{pmatrix}.$$

- Holm, Hochberg procedures with the use of statistics $z_{i,j}$

Quality of statistical procedures.

- Let $S = (s_{i,j})$, $Q = (q_{i,j})$, $S, Q \in \mathcal{G}$ - set of all adjacency matrices.
- H_S-hypothesis that threshold graph has adjacency matrix $S, S \in \mathcal{G}$.
- d_Q-decision, that threshold graph has adjacency matrix $Q, Q \in \mathcal{G}$.
- $w(H_S; d_Q) = w(S, Q)$ - loss from the decision d_Q when the hypothesis H_S is true, $w(S, S) = 0, S \in \mathcal{G}$.
- Risk function of statistical procedure $\delta(x)$ is defined by

$$Risk(S; \delta) = \sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q / H_S), \quad S \in \mathcal{G}$$

$P(\delta(x) = d_Q / H_S)$ - the probability that decision d_Q is taken while the true decision is d_S. Risk function reflects a quality of statistical procedure $\delta(x)$.

Quality of statistical procedures. Unbiasedness.

Decision function $\delta(x)$ is said to be w-unbiased if for all θ, θ'

$$E_\theta w(\theta', \delta(X)) \geq E_\theta w(\theta, \delta(X))$$

"δ is unbiased if on the average $\delta(X)$ comes closer to the correct decision than to any wrong one" (Lehmann, Romano, 2005)
In our case it can be written as

$$\sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q / H_S) \leq \sum_{Q \in \mathcal{G}} w(S', Q) P(\delta(x) = d_Q / H_S),$$

$\forall S, S' \in \mathcal{G}$

Loss function (Lehmann)

For threshold graph identification problem it is natural to consider loss functions which are additive.

$a_{i,j}$ - individual loss from false inclusion of edge (i,j) in threshold graph.

$b_{i,j}$- individual loss from false non inclusion of the edge (i,j).

Let

$$l_{i,j}(S,Q) = \begin{cases} a_{i,j}, & \text{if } \; s_{i,j} = 0, q_{i,j} = 1, \\ b_{i,j}, & \text{if } \; s_{i,j} = 1, q_{i,j} = 0, \\ 0, & \text{else} \end{cases}$$

For additive loss function one has:

$$w(S,Q) = \sum_{i=1}^{N} \sum_{j=1}^{N} l_{i,j} = \sum_{\{i,j: s_{i,j}=0; q_{i,j}=1\}} a_{i,j} + \sum_{\{i,j: s_{i,j}=1; q_{i,j}=0\}} b_{i,j}$$

Then

$$Risk(S;\delta) = \sum_{i=1}^{N} \sum_{j=1}^{N} risk(s_{i,j}, \varphi^{\delta}_{i,j}(x))$$

Optimality of δ^P

Theorem 1: Let loss function w be additive, individual test statistics $t_{i,j}$ depends only on observations $X_i(t), X_j(t)$ and vector $X = (X_1, \ldots, X_N)$ has a multivariate normal distribution. Then for single step statistical procedure δ^P for threshold graph identification in Pearson correlation network one has $Risk(S, \delta^P) \leq Risk(S, \delta)$ for any adjacency matrix S and any $w-$unbiased δ.

Optimality is proved in Koldanov A.P., Koldanov P.A., Kalyagin V.A., Pardalos P.M. Statistical Procedures for the Market Graph Construction, Computational Statistics and Data Analysis, v.68, pp.17-29 (2013).

Individual hypotheses

$$h_{i,j}: \quad \gamma_{i,j}^P \leq \gamma_0^P, \text{ vs } k_{i,j}: \quad \gamma_{i,j}^P > \gamma_0^P$$

Assumption of normality can not be removed.

Sensitivity to distribution. Robustness.

Market network. Normality is not all time observed. Heavy tails distributions. Multivariate Student distribution: an example of heavy tails distributions. Does δ^P work for threshold graph identification? Numerical experiments:

1. We consider the real-world data from USA stock market.
2. We calculate correlation matrix Σ by this data and consider the matrix Σ as true matrix.
3. We simulate n observation using the mixture distribution. The mixture distribution is constructed as follow - random vector $R = (R_1, \ldots, R_N)$ takes value from $N(0, \Sigma)$ with probability ν and from $t_3(0, \Sigma)$ with probability $1 - \nu$.
4. We estimate the matrix Σ using estimations of Pearson correlations $\hat{\rho}_{i,j}$.
5. We construct the sample market (threshold) graph and compare it to the true market graph.

Sensitivity to distribution. Robustness.

The model is the mixture distribution consisting of multivariate normal distribution and multivariate Student distribution with 3 degree of freedom.

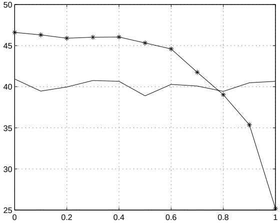

Figure: Risk function for threshold graph, $\rho_0 = 0.64$, $n = 400$, star line - δ^P

Sign similarity network

Sign similarity network: measure of association is
$\gamma_{i,j}^{Sg} = p^{i,j} = P((X_i - E(X_i))(X_j - E(X_j) > 0)$.
Individual tests

- Individual hypotheses: $h_{ij} : p^{i,j} \leq p_0$ vs $k_{ij} : p^{i,j} > p_0$
- $\varphi_{i,j}^{Sg} = \begin{cases} 0, & v_{i,j} \leq c_{i,j} \\ 1, & v_{i,j} > c_{i,j} \end{cases}$,
  $v_{i,j} = \sum_{t=1}^{n} u^{i,j}(t)$,
  $u^{i,j}(t) = \begin{cases} 1, & sign(x_i(t)) = sign(x_j(t)) \\ 0, & \text{else} \end{cases}$
  $c_{i,j}$ is defined from equation: $\sum_{k=c_{i,j}}^{n} \frac{n!}{k!(n-k)!}(p_0)^k(1-p_0)^{n-k} \leq \alpha$

Sign similarity network

- Multiple decision single step (Bonferroni type) procedure

$$\delta^{Sg}(x) = \begin{pmatrix} 1, & \varphi^{Sg}_{1,2}(x), & \ldots, & \varphi^{Sg}_{1,N}(x) \\ \varphi^{Sg}_{2,1}(x), & 1, & \ldots, & \varphi^{Sg}_{2,N}(x) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi^{Sg}_{N,1}(x), & \varphi^{Sg}_{N,2}(x), & \ldots, & 1 \end{pmatrix}.$$

- Holm, Hochberg procedures with the use of statistics $v_{i,j}$

Optimality of δ^{Sg}

Theorem 2: Let loss function w be additive, individual test statistics $t_{i,j}$ depends only on $u^{i,j}(t)$, $E(X_i) = 0, \forall i = 1, \dots, N$ and distribution of vector $X = (X_1, \dots, X_N)$ satisfy the symmetry condition below.
Then for single step statistical procedure δ^{Sg} for threshold graph identification in sign similarity network one has $Risk(S, \delta^{Sg}) \leq Risk(S, \delta)$ for any adjacency matrix S and any $w-$unbiased δ.
Symmetry condition:

$$p_{11}^{ij} = p_{00}^{ij}, \quad p_{10}^{ij} = p_{01}^{ij}, \quad \forall i, j$$

where $p_{11}^{ij} = P(X_i > 0, X_j > 0)$; $p_{00}^{ij} = P(X_i \leq 0, X_j \leq 0)$
$p_{01}^{ij} = P(X_i \leq 0, X_j > 0)$; $p_{10}^{ij} = P(X_i > 0, X_j \leq 0)$

Robustness of δ^{Sg}

Symmetry conditions are satisfied for the class of elliptically contoured distributions (ECD). Density function for ECD:

$$f(x) = |\Lambda|^{-\frac{1}{2}} g\{(x-\mu)'\Lambda^{-1}(x-\mu)\}$$

where Λ is positive definite matrix, $g(x) \geq 0$, and $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(y'y) dy_1 \dots dy_N = 1$.

Theorem 3: Let loss function w be additive and r.v. $X = (X_1, \dots, X_N)$ has a multivariate ECD with $\mu = 0$. Then conditional risk of single step statistical procedure δ^{Sg} for threshold graph identification in sign similarity network does not depend on g.

Role of measure of association

Good news: we have a distribution free (robust) multiple testing statistical procedure in sign similarity network.

Question: can we do it in Pearson correlation network?

Answer is "'YES"

Theorem 4: If $X = (X_1, \ldots, X_N)$ has a multivariate ECD with $\mu = 0$ then there is one to one correspondence between threshold graphs in Pearson correlation and sign similarity networks given by

$$p^{i,j} = \frac{2}{\pi} \arcsin \rho_{i,j}$$

,

$$\rho_{i,j} = \sin \frac{\pi}{2} p^{i,j}$$

Large scale networks

- New phenomenons are observed. Properties of multiple testing statistical procedures for threshold graph identification depend on concentration of correlations
- Quality of Holm and Hochberg procedures in Pearson correlations network essentially depend on distribution.
- Holm and Hochberg procedures, based on statistics $v_{i,j}$ (sign similarity network) are distribution free statistical procedures for threshold graph identification.

THANK YOU FOR YOUR ATTENTION!